Une expérience d'annotation à large échelle : le projet OTIM

Philippe Blache, Roxane Bertrand, Brigitte Bigi, Robert Espesser, Mathilde Guardiola, Stéphane Rauzy

Une expérience d'annotation à large échelle : le projet OTIM

P. Blache, R. Bertrand, B. Bigi, R. Espesser, M. Guardiola & S. Rauzy
LPL-CNRS, Université de Provence

Nous proposons dans cette présentation de faire le point sur une opération d'annotation de grande envergure conduite dans le cadre du projet OTIM[1].

Nous avons dans le cadre de ce projet constitué un grand corpus audio-visuel de parole spontanée comprenant 8 heures de dialogues (soit 102.457 mots correspondant à 6.611 formes différentes) totalement transcrit, aligné et richement annoté pour l'ensemble des domaines et des modalités. Nous avons donc été confrontés aux principaux problèmes posés par l'annotation de ce type de ressource. Cette présentation décrit les recommandations et les techniques que nous avons utilisées pour parvenir à nos fins.

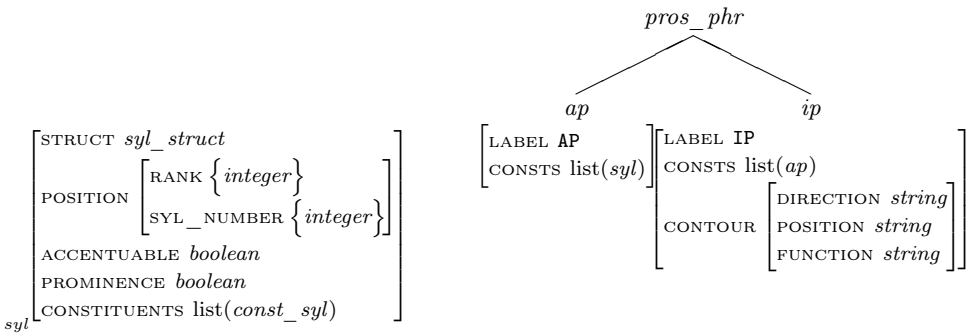
Transcription De façon à préserver les données, la transcription répond aux exigences établies de longue date par le GARS de façon à préserver les données. Nous avons par ailleurs normalisé la transcription d'un certain nombre de phénomènes de façon à prendre en compte les phénomènes spécifiques de l'oral conversationnel (élisions, assimilations, réalisations phonétiques particulières, etc.) et permettre la génération de la transcription phonétique et son alignement avec le signal . Ces conventions de transcriptions forment ce que nous appelons la transcription orthographique enrichie (TOE). A partir de la TOE il est de plus possible de générer automatiquement la transcription phonétique, codée en SAMPA.

Alignement Notre objectif dans le cadre de ce projet est de disposer d'annotations précises pour tous les domaines, de la phonétique au discours. Nous avons donc besoin d'un alignement au niveau des phonèmes. Nous utilisons pour cela l'aligneur du LORIA permettant de fournir la localisation temporelle de chaque phonème sur le signal. À partir de cet alignement et de la TOE, un module permet d'établir la correspondance entre token phonétique et token orthographique.

Anonymisation La procédure d'anonymisation consiste à éliminer des canaux audio les informations personnelles des locuteurs (noms, lieu de résidence, etc.). Un repérage automatique des noms propres a été réalisé, puis corrigé manuellement de façon à n'éliminer que les données personnelles et à vérifier que l'alignement mot/audio est correct. La transcription est anonymisée en remplaçant ces noms propres par une étiquette unique pour chaque entité (constituant une trace pour l'annotation des phénomènes de référence). Un outil d'anonymisation est ensuite appliqué sur chaque piste, bruitant le signal tout en préservant la F0 et l'intensité.

Schéma d'encodage Nous avons dans le cadre du projet proposé de standardiser la conception du schéma d'encodage en élaborant un schéma abstrait. Celui-ci consiste à décrire chaque domaine d'information sous la forme d'une structure de trait typée. Par exemple, la structure de traits suivante représente les informations associées à la syllabe ou aux syntagmes prosodiques.

[1] *Outils pour le Traitement de l'Information Multimodale*, http://www.lpl-aix.fr/~otim



Ce modèle abstrait offre plusieurs avantages. Tout d'abord, il permet de proposer une représentation globale et homogène de l'information, permettant de décrire précisément les traits, leurs types et leur organisation. De plus, ce modèle permet de générer automatiquement un schéma XML dans lequel les annotations seront encodées. Il devient donc possible de produire un schéma d'encodage générique et réutilisable.

Annotations Le travail d'annotation a porté sur un grand nombre de domaines : phonétique, prosodie, phonologie, syntaxe, discours et gestes. La phonétique contient ainsi toutes les informations y compris articulatoires sur les phonèmes. L'annotation prosodique manuelle comporte quant à elle des informations variées comme les constituants, les contours, la proéminence, etc. La structure tonale, suivant le modèle INTSINT, est quant à elle générée automatiquement. De même, la structuration syllabique est obtenue par un syllabeur que nous avons développé. Notre analyseur morpho-syntaxique, permet de déterminer les catégories à partir desquelles une analyse superficielle en chunks est effectuée de même qu'une segmentation en pseudo-phrases. Une annotation manuelle des constructions détachées a par ailleurs été réalisée, de même que pour les disfluences (avec l'aide d'un système automatique de repérage). Les annotations sur le discours portent quant à elles sur différents phénomènes comme l'humour, les backchannels, le discours rapporté, les phases de narration, etc. Les gestes enfin ont été décrits manuellement. Pour certains phénomènes (notamment tous ceux obtenus automatiquement), la totalité du corpus a été annoté, tandis que d'autres domaines nécessitent encore un travail manuel considérable (c'est le cas des gestes qui ont été annotés sur un dialogue d'une heure).

Outils Nous avons utilisé pour les annotations manuelles Praat et Anvil. Les traitements automatiques ont quant à eux été effectués par des systèmes développés au LPL. Un travail important a été entrepris concernant l'interopérabilité. Tout d'abord, toutes les annotations automatiques génèrent un format respectant le schéma XML obtenu à partir de la description en structure de traits typés décrite plus haut. Le format obtenu peut ainsi servir de langage pivot pouvant être utilisé par les différents éditeurs. Nous avons ainsi développé un utilitaire permettant d'éditer sous Praat (en lecture et écriture) les données encodées selon ce schéma XML. Ce même utilitaire est en cours de développement pour Anvil. Nous obtenons par conséquent une possibilité d'interopérabilité entre ces deux éditeurs.

Patrimonialisation, diffusion Les données du CID et leurs annotations ont été déposées sur le centre de ressource numérique CRDO-Aix. La totalité du corpus (signal audio et vidéo), sa transcription orthographique et phonétique sont disponibles en libre accès. Les enrichissements sont eux disponibles gratuitement sous condition.

References

[Allwood05] Allwood J., L. Cerrato, L. Dybkjaer and al. (2005) The MUMIN Multimodal Coding Scheme, NorFA yearbook 2005

[Bertrand08] Bertrand, R., Blache, P., Espesser, R., Ferré, G., Meunier, C., Priego-Valverde, B., Rauzy, S. (2008) "Le CID - Corpus of Interactional Data - Annotation et Exploitation Multimodale de Parole Conversationnelle", in revue *Traitement Automatique des Langues*, 49:3.

[Bird00] Bird S., Day D., Garofolo J., Henderson J., Laprun C. & Liberman M. (2000) "ATLAS : A Flexible and Extensible Architecture for Linguistic Annotation", in procs of *LREC00*

[Bird01] Bird S., M. Liberman (2001) "A formal framework for linguistic annotation" *Speech Communication*, 33:1-2

[Blache09] Blache P., R. Bertrand, and G. Ferré (2009) "Creating and Exploiting Multimodal Annotated Corpora: The ToMA Project". in Kipp, Martin, Paggio and Heylen (eds.) *Multimodal Corpora: From Models of Natural Interaction to Systems and Applications*, LNAI 5509, Springer.

[Blache10] Blache P. et al. (2010) "Multimodal Annotation of Conversational Data", in proceedings of *LAW-IV - The Linguistic Annotation Workshop*

[Boersma09] Boersma P. & D. Weenink (2009) *Praat*: doing phonetics by computer, http://www.praat.org/

[Dipper07] Dipper S., M. Goetze and S. Skopeteas (eds.) (2007) *Information Structure in Cross-Linguistic Corpora: Annotation Guidelines for Phonology, Morphology, Syntax, Semantics and Information Structure*, Working Papers of the SFB 632, 7:07

[Dybkjaer01] Dybkjaer L., S. Berman, M. Kipp, M. Wegener Olsen, V. Pirrelli, N .Reithinger, C. Soria (2001) "Survey of Existing Tools, Standards and User Needs for Annotation of Natural Interaction and Multimodal Data", *ISLE Natural Interactivity and Multimodality Working Group Deliverable D11.1*

[Ide07] Ide N. and K. Suderman (2007) "GrAF: A Graph-based Format for Linguistic Annotations" in proceedings of the *Linguistic Annotation Workshop* (LAW-07)

[Ide09] Ide N. and Suderman K. (2009) Bridging the Gaps: Interoperability for GrAF, GATE, and UIMA. Proceedings of the Third Linguistic Annotation Workshop, held in conjunction with ACL 2009, Singapore.

[Illina04] Illina I., D. Fohr, O. Mella, C. Cerisara (2004) The Automatic News Transcription System: ANTS some Real Time experiments, Proceedings of the 8th International Conference on Spoken Language Processing, Jeju, Corée du Sud,

[Kipp01] Kipp M. (2001) "Anvil-a generic annotation tool for multimodal dialogue" in procs of 7th European Conference on Speech Communication and Technology

[McNeill05] McNeill, D. (2005) *Gesture and Thought*, The University of Chicago Press.

[Rodriguez07] Rodriguez K., Stefan, K. J., Dipper, S., Goetze, M., Poesio, M., Riccardi, G., Raymond, C., Wisniewska, J. (2007) "Standoff Coordination for Multi-Tool Annotation in a Dialogue Corpus", in procs of the *Linguistic Annotation Workshop at the ACL'07 (LAW-07)*